Audiovizuális beszédfelismerés

Czap László

Miskolci Egyetem, Villamosmérnöki Intézet, Automatizálási Tanszék
3515 Miskolc, Egyetemváros
czap@mazsola.iit.uni-miskolc.hu

Abstract. Az emberi beszédértés bimodális természetű: az akusztikus és vizuális jelet zseniálisan kombináljuk a maximális érthetőség érdekében. Különösen zajos környezetben segíti a beszéd jobb megértését a vizuális jel. A szájról olvasás feladatát próbálom gépi úton megvalósítani. Az audiovizuális beszédfelismerés fő kérdései, hogy mely jellemzők hordozzák a lényegi vizuális információt, és hogy ezek hogyan nyerhetők ki a képből. A geometriai és pixel bázisú lényegkiemelést a folyamatos beszédfelismerés szempontjai szerint még nem hasonlították össze. Arra a kérdésre is választ kerestem, hogy eséllyel léphet-e fel a diádok vetélytársaként a félszótag, mint a felismerés alapegysége.

1. Bevezetés

Az emberi kommunikáció multimodális természetű. A multimodalitást értelmezhetjük úgy, hogy a kommunikációban több érzékszervünk vesz részt. A hallás mellett a látás a legfontosabb információforrásunk. Bernsen [1] összekapcsolja a modalitást a médiummal, mint az információ valamely formájának fizikai hordozójával. A média rokonítható az érzékszervekkel, amelyekre hat. A grafikus médium pl.: a látással, az akusztikus médium a hallással társítható. A technika fejlődésével az ember-gép kommunikációban is egyre több modalitás juthat szerephez.

Cikkem arról a munkáról számol be, amelynek keretében a vizuális modalitás által hordozott információval egészítem ki az akusztikus modalitás elemzését, a szájról olvasást próbálom gépi úton megvalósítani. Massaro [3] kísérletekkel igazolta, hogy a modalitásokat egymás kiegészítésére használjuk. Ha a hang gyenge minőségű, vagy hallássérült a megfigyelő, jobban hagyatkozik a szájról olvasásra. „Jobban hallom a TV-t, ha felteszem a szemüvegem." Az emberi beszédértést meg sem közelítő gépi felismerőket hasonlíthatjuk a környezet vagy képességei által korlátozott emberi felfogóhoz abban a tekintetben, hogy a kiegészítő vizuális jel a gépi beszédfelismerők felismerési hatékonyságát is javíthatja, különösen zajos környezetben.

2. Az emberi bimodális beszédfelismerés analízise

Napjainkban a gépi beszédfelismerés szédületes ütemű fejlődésének vagyunk tanúi. Ezek megbízhatósága azonban jelentősen romlik zajos beszéd esetén. Egyik kitörési pont lehet a vizuális jel felhasználása a beszéd felismeréséhez. A gépi szájról olvasás tervezéséhez célszerű ismerni, hogy az emberi kommunikációban az arc mely részei mennyire segítik a beszéd felismerését. Ezeket a vizsgálatokat zajos beszéddel végezhetjük el, hiszen a jó minőségű beszéd tökéletesen érthető, a vizuális támogatás hatása nem mérhető.

Ebben a fejezetben arra keressük a választ, hogy az arc mely részei hordozzák a legtöbb vizuális információt a beszédfelismeréshez. Érthetőség vizsgálatot végeztünk zajos beszéddel úgy, hogy az arc egyes részeinek maszkolásával a beszélő arcának csak részletei voltak láthatók.

Várakozásaink szerint az ajakforma lényeges vizuális jellemző. Fontos kérdés, hogy a nyelv és a fogak láthatósága számottevő javulást okoz-e, érdemes-e a lényegkiemelésnél erőfeszítéseket tenni a leírásukra. Az arc egyéb részei is hozzájárulnak a beszédfelismerési eredmények további javulásához, ha a beszélő egész arca látható? Ezekre a kérdésekre kerestük a választ szubjektív teszt segítségével.

2. 1. Érthetőség vizsgálat

Az arc különböző részletei által hordozott vizuális támogatás mérésére érthetőség vizsgálatot végeztünk. Mássalhangzó felismeréshez V_1CV_1 szavak (pl.: eke, ata) középső mássalhangzóját kellett megjelölni. Magánhangzó felismerésekor C_1VC_1 szavak (pl.: lol, tet) középső magánhangzóját kellett kitalálni.

A teszteket 78, fonetikai ismeretekkel nem rendelkező egyetemi hallgató töltötte ki. A 10-15 fős csoport egy közös TV készüléken nézte a videó jelet és egy közös hangszóróból hallotta a hangot, kétszer egymás után. A válaszra korlátozott idő – kb. 3 másodperc – állt rendelkezésre.

A vizuális jel az alábbiak valamelyike lehetett:

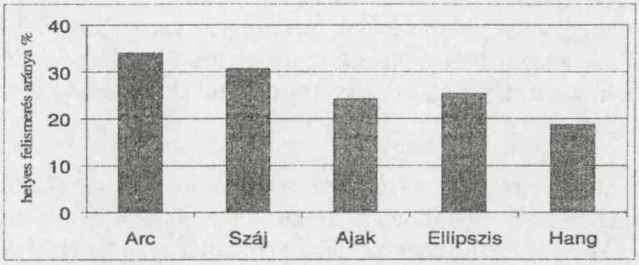
- a beszélő arca
- a beszélő szája (ajkak, fogak, nyelv)
- a beszélő ajkai
- az ajkak méreteit utánzó ellipszis

Az akusztikus jel zajos beszéd volt. A magánhangzó felismerési kísérleteket -18 dB pillanatnyi jel-zaj viszony mellett végeztük. A mássalhangzó felismerési vizsgálatoknál a jel-zaj viszony -6 dB volt. A pillanatnyi jel-zaj viszonyhoz szükséges zaj amplitúdót 5 ms-onként állítottuk be.

Egyes kísérleteknél csak akusztikus jel volt jelen (sötét képernyő), máskor hang nélkül, csak a kép alapján próbáltuk a magánhangzót vagy a mássalhangzót felismerni.

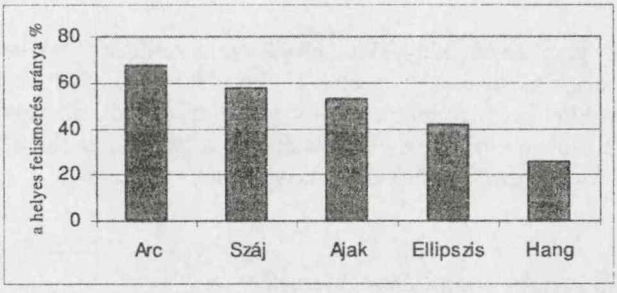
Az eredmények 11 623 válasz kiértékelése alapján születtek, ezek közül 9 625 a mássalhangzó, 1 998 a magánhangzó felismerését szolgálta. Egy hallgató egy hang megjelölésével adott egy választ.

Az 1. ábra a mássalhangzó érthetőség vizsgálat eredményét mutatja audiovizuális jel esetén.



1. ábra. Mássalhangzók felismerési aránya. A hang pillanatnyi jel-zaj viszonya: -6 dB, a képen az arc részletei, vagy a száj vízszintes és függőleges méretével megegyező kis- és nagy tengelyű ellipszis volt látható. Az Hang megnevezésű oszlop esetében a képen csak a minta sorszáma volt látható.

A 2. ábrán a magánhangzók felismerési eredményeit láthatjuk.



2. ábra. A magánhangzók felismerési arányai. A pillanatnyi jel-zaj viszony -18 dB.

Várakozásunkkal egyezően, minél többet látunk a beszélő arcából, annál jobban segíti a kép a beszéd felismerését. Mivel jelentéssel nem bíró szavakról van szó, az arckifejezés nem növelhette az érthetőséget az egész arc megmutatása esetén sem. A javulás a száj (ajkak, fogak, nyelv) figyeléséhez képest inkább annak tulajdonítható, hogy az arc redői kiemelik a szájmozgást, segítik az artikuláció pontosabb követését.

3. A vizuális lényegkiemelés

A szubjektív tesztek bizonyították, hogy az ellipszis méretét leíró ajakszélesség (a) és ajaknyílás (b) az ajkak láthatóságához hasonló felismerési eredményeket hozott. A képmomentumokból származtatható intenzitás faktor (k) - amely valójában a szájnyílás átlagos világossága - szolgált a magánhangzó és mássalhangzó felismerési kísérle-

teknél a nyelv és a fogak láthatóságának jellemzésére. A *k* intenzitás faktor a hátul képzett hangoknál a legkisebb (pl.: k, u). Közepes értékű, ha elül képzett hangoknál a nyelv látható (pl.: e, i). Legnagyobb a *k* értéke, ha a fogakat látjuk a szájnyílásban (pl.: s, cs).

A vizuális jellemzők kinyerésére kifejlesztett eljárások közös jellemzője, hogy a száj belső és külső kontúrjának követését követelik meg. [2, 5] Ezek a módszerek rendkívül számításigényesek, ezért lassúak. A számítási kapacitás növekedésével ez a probléma enyhül, de a legújabb módszerek sem elég megbízhatók. Az általam javasolt eljárás nem igényli a száj kontúrjának követését. A feldolgozásra kijelölt terület képi hasonlóságán alapul.

A számítógépek sebességének és tárkapacitásának további növekedésével a geometriai alapú rendszerek mellett előtérbe került a pixel bázisú feldolgozás. Ebben az esetben a száj és környezete, de akár az egész kép minden pontja részt vehet az elemzésben. A pixel bázisú feldolgozás előnye, hogy az ajkak környezetének feldolgozásával a szájmozgást kiemelő redőzet is a felismerés szolgálatába állítható. A teljes arc feldolgozása esetén a gesztusok figyelembe vételére is lehetőség nyílik. A geometriai alapú feldolgozás lehetővé teszi az artikuláció elemzését, a hangképzés statikus és dinamikus jellemzőinek mérését. A pixel alapú feldolgozás ezeket a lehetőségeket nem kínálja. Hátránya a pixel bázisú feldolgozásnak, hogy érzékenyebb a megvilágítás változásaira és személyfüggő felismeréshez használható.

A geometriai és pixel bázisú lényegkiemelés összehasonlítását csak igen kis méretű adatbázison, szó alapú elemzéssel végezték el. Feladatomnak tekintem a folyamatos audiovizuális beszédfelismerés szempontjait szem előtt tartó összehasonlító vizsgálat elvégzését. Az artikuláció dinamikus vizsgálatát csak a geometriai alapú elemzés szolgálja, ezért ennek kidolgozását elsőrendű fontosságúnak tartottam.

3. 1. A geometriai alapú vizuális lényegkiemelés

A jellegzetes képeken a szájsarkak elmozdulás vektorok vagy manuális segítség alapján kijelölhetők. Az ajkak méretét reprezentáló ajakszélesség (*a*) és ajaknyílás (*b*) ezek alapján meghatározható. Az ajkak belső és külső kontúrjainak követése igen nehézkes, ezért olyan módszer kifejlesztésére törekedtem, amely ezt nem igényli. A következő bekezdésben tárgyalt prototípus alakzatok jellemzőinek kialakításánál megengedhetőnek tartom a kézi beavatkozást, mivel csak a prototípus alakzatokra kell meghatározni őket. A szájnyílás belső területét a belső szájsarkak és az ajaknyílás felső illetve alsó széle közé rajzolt parabolák által határolt területtel közelítem, amelyre a *k* intenzitás faktor meghatározható.

Az adatbázisban szereplő videó anyag képkockáin kijelölt feldolgozandó terület jellegzetes ajakformáinak a különböző nyelvállásokat és a fogak eltérő láthatóságát figyelembe vevő prototípus alakzatok kiválasztásával artikulációs könyvtárt hoztam létre. Ezeken a képeken elvégeztem – a jellegzetes pontok esetleg manuális kijelölésével – a lényegkiemelést. Az adatbázis összes képének feldolgozásával meghatározva

képkockánként a hasonlóság mértékét, a legkevésbé hasonló alakzatokat felvettem az artikulációs könyvtárba. A műveletet addig ismételtem, amíg a legkevésbé hasonló alakzatok jellemzői bele nem simultak a környezetükbe. A prototípus alakzatok kiválogatása után 88 alakzat képviselte a jellegzetes képeket.

A módszer számos előnnyel jár az ismert eljárásokkal összehasonlítva:

- mérsékelt számításigényű, valós időben is elvégezhető a mai PC-ken
- a száj környezetét is figyelembe veszi, feldolgozási területe a geometriai alapú és a pixel bázisú módszerrel feldolgozott terület között helyezkedik el
- tetszőleges jellemzőket választhatunk a lényegkiemelésre, ezeket csak a kiválasztott képekre kell meghatározni, manuális támogatás is adható
- nem igényli a száj sem külső, sem belső kontúrjának meghatározását
- fekete-fehér képeken elvégezhető
- a vektorkvantáláson alapuló feldolgozás esetén közvetlen bemenetként szolgálhat

Hátránya, hogy beszélőfüggetlen feladathoz az artikulációs könyvtár bővítésére van szükség, ami a feldolgozási idő növekedéséhez vezet. A kutatás jelenlegi fázisában a beszélőfüggetlen audiovizuális beszédfelismerés nem tűzhető ki reális célként. Külön kutatási terület lehet az artikuláció személyfüggősége, a vizuális és az akusztikus jellemzők összefüggése.

3. 2. A pixel bázisú lényegkiemelés

A pixel bázisú lényegkiemelés az ajkak környezetének kijelölésével, rendszerint a képpontok számának decimációval történő redukálása után elvégzett transzformációt jelenti. A transzformációk a képsíkból a síkfrekvencia tartományba konvertálják a képeket. Erre a célra a diszkrét koszinusz transzformációt választottam, amelynek előnye a Fourier transzformációval szemben, hogy valós függvényeket valós függvényekbe konvertál. [4] A száj környezetének kijelölése a geometriai alapú feldolgozáshoz megtörtént, a sorok és oszlopok számát decimálással harmadolva 27x23 pontos képet kapunk. A transzformált jelből vízszintes és függőleges irányú síkfüggvényekből a 8-8 legkisebb síkfrekvenciájú 64 bázisfüggvény együtthatóiból válogattam a vizuális jellemzőket. A nagy síkfrekvenciájú komponensek a textúrát képviselik.

4. A felismerés alapegysége

Másik lényeges kérdés, amelyre a választ keresem cikkemben, hogy mit célszerű a gépi beszédfelismerés során a beszéd felismerendő nyelvi egységének tekinteni. Agglutináló nyelvek esetében a szóalapú feldolgozás folyamatos beszéd felismerésére nem alkalmas. Nincs általánosan elfogadott becslés a magyar nyelvben előforduló szóalakok számára vonatkozóan, annyi azonban bizonyos, hogy kezelhetetlen mennyiségről

van szó. Nyilvánvaló, hogy a fonéma szintű felismerés a hangok egymásra hatása miatt nem lehetséges. A szónál rövidebb, a fonémánál hosszabb alakzatokat célszerű választani felismerendő nyelvi egységként. A diád alapú és a Vicsi Klára [6] által javasolt félszótag alapú gépi beszédfelismerést kívánom összevetni. Összehasonlító elemzést végeztem a diád és félszótag alapú gépi beszédfelismerés tekintetében.

Vicsi Klára elemzései szerint a magyar nyelvű szövegek részleges lefedéséhez a félszótagokból kell a legszűkebb készletet figyelembe venni. A félszótag azért is ígéretes jelöltnek tűnik, mert általában hosszabb a diádnál, és hosszabb elemeket könnyebb megkülönböztetni egymástól. Hátránya, hogy a kezdő és záró félszótag csak a magánhangzónál illeszthető, az így képzett szótaghatárokon a hangok egymásra hatását nem tudja figyelembe venni.

A félszótagok egyik vetélytársa a diád lehet. A diád előnye, hogy mindkét végén illeszthető, figyelembe tudja venni a koartikulációs hatásokat. További előnye, hogy a teljes lefedéshez kevesebb elemre van szükség, mint a félszótagok esetében. Hátránya, hogy átlagos időtartama a félszótagokénál rövidebb.

4. 1. Az audiovizuális és az akusztikus adatbázis

Magyar nyelven nem áll rendelkezésre audiovizuális beszéd adatbázis, ezért a szerző bemondásával, házi videó berendezés és közönséges PC mikrofon felhasználásával felvett hang- és videó anyagon folyt a tanítás és tesztelés. A felvételek egyféle beállítással, az ülő helyzetben természetes fejmozgás mellett, speciális világítási előírások nélkül készültek. Az általánosabb alkalmazhatóság érdekében a színinformációt nem akartam felhasználni, ezért a képek feldolgozása a színes képek intenzitás képpé alakításával kezdődött. A felvétel körülményei normál irodai környezetnek felelnek meg. A hang a szobában működő, ventilátor zajt termelő PC-n került rögzítésre. Az utcáról beszűrődő zaj mellett a számítógép tápegysége által okozott zaj jelentős mértékű.

Az adatbázis felvételénél személyfüggő elemzést tűztem ki célul. Az audiovizuális adatbázis a számok és dátumok félszótag készletén alapul, a tanításra 486 szótag és 79 szó szolgált, a tesztelésre 35 szófüzért használtam. Az audiovizuális adatbázis a képkockák félképekre bontásával másodpercenként 50 képet tartalmaz, a szinkronitás végett az akusztikus jel is 20 ms-os lépésközzel kerül feldolgozásra. Az akusztikus lényegkiemelés 12 MFCC együtthatót tartalmaz, a mintavételi frekvencia 22 050 Hz.

Annak megbízhatóbb eldöntésére, hogy a felismerés alapegységeként a félszótag vagy a diád az alkalmasabb választás, egy nagyobb akusztikus adatbázist hoztam létre, saját bemondással 8000 szó szolgálta a tanítást, 1400 szó a tesztelést. A szavakban szereplő 121 kezdő félszótag kumulatív gyakorisága 70,2%, a 83 záró félszótag kumulatív gyakorisága 80,7%, a kiválasztott félszótagok a szótagok 60,1%-át, a szavak 32,6%-át fedték le. A statisztikai elemzés 1 996 589 szóból álló klasszikus és modern prózát dolgozott fel, amely 4 238 066 szótagot tartalmazott.

A diádokon alapuló elemzés ugyanezen az adatbázison történt. Az audiovizuális adatbázis a diádok teljes készletének mintegy 20%-át, az akusztikus adatbázis a diádok 50%-át fedte le.

5. A beszédfelismerési eredmények

A beszéd felismerését az audiovizuális adatbázison bimodális és unimodális gerjesztéssel is megvizsgáltam, mind félszótag, mind diád alapú elemzéssel.

5. 1. Az audiovizuális adatbázis felismerési eredményei

A geometriai bázisú vizuális lényegkiemelés félszótag alapú felismerési kísérletekben a hibák ötödét kiküszöbölte. A diád alapú beszédfelismerési kísérletben a geometriai bázisú lényegkiemelés vizuális kiegészítő jele alig volt képes javítani az akusztikus felismerési eredményeken. A pixel bázisú vizuális lényegkiemelés eredményeképpen a felismerési hibákat mintegy felére sikerült csökkenteni.

1. Táblázat. A helyes felismerés arányai az audiovizuális adatbázison

	Félszótag	Diád/Geometriai	Diád/Pixel
Akusztikus	69,1 %	88,7	88,7
Audiovizuális	75,9 %	89,8	94,2

5. 2. Az akusztikus adatbázis felismerési eredményei

Az akusztikus adatbázison végzett kísérletek is igen lényeges különbséget mutattak a félszótag és diád alapú felismerési eredményekben. A pusztán akusztikus gerjesztés esetében a félszótag alapú felismerés hibáit felezte a szótagok kapcsolódását kifejező *kötött félszótag* alapú felismerés bevezetése. A diád alapú felismerés hibáinak száma mintegy ötöde a kötött félszótag alapú felismerési hibáknak.

2. Táblázat. A helyes felismerés arányai az akusztikus adatbázison

	Félszótag	Kötött félszótag	Diád
Akusztikus	59,2 %	79,8 %	95,9 %

A kötött félszótag értelmezése: A koartikulációs hatások figyelembe vétele céljából a szótaghatárokon a félszótagokat diádokkal illesztve, (diád - kezdő félszótag - záró félszótag) sorozatok alkotják a láncot, a végén diáddal lezárva. A szótagok ilyen illesztése felére csökkentette a hibák számát, de a diád alapú felismerés eredményeit nem tudja megközelíteni.

6. Összefoglalás

Az audiovizuális beszédfelismerési kísérletek megmutatták, hogy a vizuális jel lényeges kiegészítő információval szolgálhat a beszéd felismeréséhez. A geometriai alapú vizuális lényegkiemelést sikerült az ajakkontúrok követése nélkül megoldani. A geometriai bázisú lényegkiemelés eredményei alapadatokat szolgáltattak az artikuláció dinamikus jellemzőinek elemzéséhez, amelyre egy vizuális beszédszintézis projekt épült. A pixel bázisú lényegkiemelés további javulást eredményezett. A vizsgálatok másik eredménye, hogy az ígéretesnek tűnő félszótag alapú felismerés nem képes a diád alapú felismerés eredményeit megközelíteni. A felismerés alapegysége kontextusfüggő, és ezt az elemeknek ki kell fejezniük.

Irodalomjegyzék

1. N. O. Bernsen: Multimodality in Language and Speech Systems – from Theory to Design Support Tool. in Multimodality in Language and Speech Systems. Kluwer Academic Publishers, Dordrecht/Boston/London 2002.
2. Luettin, J., Thacker, N.A., and Beet, S.W. (1996). Speechreading using shape and intensity information. Proc. International Conference on Spoken Language Processing, Philadelphia, PA, pp. 58–61.
3. Massaro, D.W. (1996). Bimodal speech perception: A progress report. In Stork, D.G. and Hennecke, M.E. (Eds.), Speechreading by Humans and Machines. Berlin, Germany: Springer, pp. 79–101.
4. Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., Vergyri, D., Sison, J., Mashari, A., and Zhou, J. (2000). Audio-Visual Speech Recognition. Final Workshop 2000 Report. Baltimore, MD: Center for Language and Speech Processing, The Johns Hopkins University.
5. Silsbee, P.L. and Bovik, A.C. (1996). Computer lipreading for improved accuracy in automatic speech recognition. IEEE Transactions on Speech and Audio Processing, 4(5):337–351.
6. Vicsi, K., Vigh, A. Text independent neural network/rule based hybrid, continuous speech recognition. EUROSPEECH'95. Madrid: pp. 2201-2204, 1995.